

**PATENT APPLICATION**

**METHODS OF MAKING HYBRID PROTEINS**

Inventor(s): Peter B. Vander Horn, a citizen of the United States of America,  
residing at 130 Trimaran Court, Foster City CA 94404

Assignee: MJ Bioworks, Inc.

Entity: Small

## METHODS OF MAKING HYBRID PROTEINS

### CROSS-REFERENCES TO RELATED APPLICATIONS

- 5 [0001] This application claims the benefit of U.S. Provisional Application No. 60/463,781, filed April 17, 2003; and U.S. Provisional Application 60/483,287, filed June 27, 2003, each of which applications is herein incorporated by reference.

### FIELD OF THE INVENTION

- 10 [0002] This invention relates to methods for the facilitated evolution of proteins.

### BACKGROUND OF THE INVENTION

- [0003] This invention provides methods of creating hybrid proteins to identify proteins with enhanced activity. A number of methods of generating hybrid sequences to enhance  
15 protein function are known (*see, e.g.*, U.S. Patent No. 6,132,970). However, these methods rely on recombinational techniques that shuffle the sequences to create new proteins. There is a need for additional techniques to facilitate identification of proteins with an enhanced function. This invention addresses that need.

### BRIEF SUMMARY OF THE INVENTION

- 20 [0004] The invention provides a method of generating polypeptides with enhanced function. The method comprises creating hybrid proteins having a common biological activity comprising the steps of: (a) creating a library of 32 or more nucleic acids encoding a plurality of hybrid protein members, wherein the members differ from a set of at least two  
25 starting proteins with corresponding amino acids, and i ) where the starting proteins are homologous proteins having greater than 60% amino acids pair-wise similarity to each other and having at least one common biological activity, ii.) where a majority of the encoding library members have a greater than 60% amino acid similarity to any of the starting proteins, and iii.) where the majority of differences between the encoded library members and the  
30 starting proteins are confined to those corresponding amino acids that differ among the

starting proteins; (b) expressing protein from at least one library member to create at least one hybrid protein; and (c) selecting at least one protein having a common biological activity of the starting proteins.

[0005] In another aspect, the invention provides a library of nucleic acids encoding a plurality of hybrid protein members, wherein the members differ from a set of at least two starting proteins with corresponding amino acids, and i) where the starting proteins are homologous proteins having greater than 60% amino acid sequence similarity to each other and having at least one common biological activity, ii) where a majority of the library members have a greater than 60% amino acid similarity to any of the parent proteins, and iii) where the majority of differences between the library members and the starting proteins are confined to those corresponding amino acids that differ among the starting proteins. In some embodiments, the parent proteins are enzymes, *e.g.*, polymerases, biosynthetic and catabolic enzymes. Parent enzymes can also be isozymes. Parent proteins can also be non-enzymatic proteins, *e.g.*, proteins that bind to another molecule, with or without allosteric effect, such as hormones, receptors, antibodies and the like. Often, the parent protein have greater than 80% amino acid similarity to each other and the majority of the library members have greater than 80% amino acid similarity to any of the starting proteins.

[0006] In another aspect, the invention provides a synthetic hybrid protein comprising greater than 60% amino acid similarity to each member of a set of at least two starting proteins, where each starting protein in the set shares greater than 60% amino acid similarity and at least one common biological activity with each member of the set, and wherein the hybrid protein: (a) shares at least one biological activity with all members of the set; (b) has a minimum of 5 amino acid residue differences from any member of the set; and (c) comprises no more than 24% of amino acids not corresponding to any member of the set.

[0007] In some embodiments, the starting parent proteins are enzymes, *e.g.* polymerases. The parents may also be isozymes.

[0008] Often, the parent proteins have greater than 80% similarity to each other and the majority of the library members have greater than 80% similarity to any of the parent proteins. hybrid protein.

[0009] In some embodiments, wherein the set of parent proteins comprises the *Pyrococcus furiosus* family B DNA polymerase (Pfu) and Deep Vent® DNA Polymerase and the

differences from any member of the set comprise at least 10 of the mismatches selected from the group indicated in Figure 8.

[0010] The invention relates to generating hybrid polypeptides that comprise alterations in less conserved regions of parent proteins and, surprisingly, provides hybrid proteins that have improvement in desired properties relative to parent proteins. A protein of the invention can be designated as a variable residue altered hybrid protein (VRAHP), as described broadly. More specifically, a VRAHP contains alterations at non-conserved positions of the parent proteins, *i.e.*, variable residues, where the altered residue is an amino acid that occurs at that position in one of the parent proteins. Such alterations are typically present throughout the protein, for example occurring in at least 1 of every 30 or 50 amino acid residues, rather than concentrated in one region of the protein.

[0011] Typically, according to the invention, each parent protein in a set of at least two parent proteins shares greater than 60% amino acid similarity and at least one common biological activity with each member of the set. A typical hybrid protein according to the present invention will comprise greater than 60% amino acid similarity to each member of a set of at least two parent proteins, and will share at least one biological activity with all members of the set.

[0012] Furthermore, the aforementioned set of parent proteins necessarily comprises a subset of invariant amino acids that are identical among all members of the set. A typical hybrid protein according to the present invention comprises at least 95% of the subset of invariant amino acids.

[0013] Finally, the set of parent proteins necessarily comprises a subset of variable amino acids that differ among at least some members of the set. A typical hybrid protein according to the present invention will comprise a minimum of 5 amino acid residue differences from any member of the set, corresponding to members of the subset of variable amino acids. A typical hybrid protein according to the present invention will also comprise a subset of at least 5 amino acid residues from among the subset of variable amino acids, where each of the at least 5 amino acids is identical to a corresponding amino acid in at least one of the members in the parental set, and each of the subset of at least 5 amino acids, in order from N-terminus to C-terminus, is identical to a corresponding amino acid from a different one of the set of parent proteins from the previous member of the subset of at least 5 variable amino

acid residues. In other words, the typical hybrid protein contains at least 5 variable amino acid residues corresponding to alternating parent proteins.

## BRIEF DESCRIPTION OF THE DRAWINGS

5 [0014] Figure 1 shows a BLOSUM62 Substitution Matrix.

[0015] Figure 2 shows a BlastP alignment of *Pyrococcus furiosus* polymerase (Pfu) (query 1) against *Pyrococcus sp* GB-D polymerase (Deep Vent®) (subject 1)

[0016] Figure 3 shows an example of assembly PCR. In this example, 100 base pair degenerate oligonucleotides are subjected to rounds of annealing and primer extension until  
10 fragments of approximately 500 base pairs are obtained. These fragment libraries are sufficiently large in size to be easily manipulated and assembled into full length clones or libraries of full length clones by conventional molecular cloning techniques.

[0017] Figure 4 shows the sequences of the parental Dut proteins and the BLASTP alignment of the parent sequences.

15 [0018] Figure 5 shows the degeneracies at the positions that differ in the parental Dut proteins. 5A.: Aligned parental sequence showing all possible codons in order of frequency of use by *E. coli*. 5B: Consensus sequence is derived by finding codons that will encode both sequences with a minimal number of degeneracies. Codons frequently used by *E. coli* are preferred. 5C: Nucleic acid degeneracies that incorporate amino acid sequences not similar  
20 (BLOSUM 62 number is <0) to either parental amino acid sequence are removed; in this example the nucleic acid encoding the thermal stable protein sequence, AAD is used instead. These are indicated in bold. The sequence of the thermal stable enzyme is also used in deciding to retain the one gap and eliminate the 2 cases where terminating codons could be incorporated in the sequence.

25 [0019] Figure 6 shows the priming and restriction sites (bold) that were added to the ends of the sequence. In two cases, codon usage was changed to add restriction sites (underlined and in italics). The amino acids encoded by the sequence are indicated below the codons.

[0020] Figure 7 shows the minimal encoding oligonucleotide sequence to be synthesized to assemble the Dut hybrid library. The DNA sequence was converted to single letter  
30 nucleotide code using standard designations and oligonucleotide sequences were selected

(below in bold). Selections were made such that minimal degeneracies exist where primers were expected to anneal to each other during assembly. In one stretch of sequence there was no region where reasonably sized annealable oligonucleotide sequences could be selected. In this example, the ClaI site (underlined) inserted in the previous step is used to assemble a full-length protein-encoding library from 2 restriction fragments.

[0021] Figure 8 shows the minimal encoding sequence used to generate oligonucleotides encoding a Pfu/ Deep Vent® Hybrid DNA polymerase as explained in example 2. The degenerate nucleotides are in parenthesis. The amino acid sequences that differ between the parent proteins (the "mismatches") are indicated. Non-parental amino acids are indicated in bold. Examples mentioned in the text are numbered.

[0022] Figure 9 shows a comparison of the polymerase to 3' exonuclease ratios for several commercially available enzymes, including the parental proteins, and isolates from the hybrid library.

[0023] Figure 10 shows the results of a comparison of hybrid and parent polymerases. The enzymes were tested for the ability to amplify bacteriophage lambda DNA amplicons of a range of sizes, given a 30 sec or 1 min extension time. The sizes of the amplicons, in kilobases, are listed across the bottom of the lanes. Twenty units of enzyme per ml were used unless otherwise specified.

[0024] Figure 11 shows a comparison of the sequences of parent and hybrid polymerase proteins.

## DETAILED DESCRIPTION OF THE INVENTION

### A. General Overview

[0025] This invention provides methods of creating hybrid proteins having a desired phenotype. In general it is often desirable to create new proteins with functions that are similar to, but altered from, the functions of known existing proteins, *e.g.*, it may be desirable to create proteins with enhanced stability, enhanced or decreased enzymatic activity towards particular substrates, enhanced or decreased affinity for particular ligands, etc. For example, a DNA polymerase enzyme may have both polymerase and exonuclease activities, and it may be useful to create new enzymes with different ratios of those two activities. This invention provides methods of producing large numbers of proteins that can be screened for desirable properties.

[0026] With the sequencing of the human, mouse, and many invertebrate and microbial genomes essentially complete, a wide variety of gene and deduced protein sequences are available. The invention therefore also provides a method of using the raw sequence information about protein variation as a source for generating useful variant proteins.

5 [0027] In particular, the invention provides methods of synthesizing a nucleic acid library that encodes hybrids of two or more parent proteins and selecting for hybrid proteins having a desired phenotype or activity. Typically, the library will comprise 32 or more hybrid proteins. The invention also provides a library of hybrid nucleic acids encoding a plurality of hybrid proteins and synthetic hybrid proteins comprising greater than 60% amino acid  
10 similarity, often greater than 60% identity, to each member of a set of at least two parent proteins.

[0028] The practice of this invention involves the construction of recombinant proteins and their expression in host cells. Molecular cloning techniques to achieve these ends are known in the art. A wide variety of cloning and *in vitro* amplification methods suitable for the  
15 construction of recombinant nucleic acids such as expression vectors are well-known to persons of skill. General texts which describe molecular biological techniques useful herein, include Sambrook & Russell, *Molecular Cloning, A Laboratory Manual* (3rd Ed, 2001) ("Sambrook"); Kriegler, *Gene Transfer and Expression: A Laboratory Manual* (1990); and *Current Protocols in Molecular Biology*, Ausubel *et al.*, eds., 1994-1999, John Wiley &  
20 Sons, Inc ("Ausubel").

[0029] Parental sequences for generating hybrid proteins for a protein of interest may be identified by various amino acid sequence comparison methods. Using these techniques, one of skill can identify conserved regions in the nucleic acids encoding the proteins of the invention to prepare appropriate oligonucleotides that can be used to generate the hybrid  
25 proteins.

[0030] Oligonucleotides can be custom-made and ordered from a variety of commercial sources known in the art. Those that are not commercially available can be chemically synthesized using a variety of chemistries, *e.g.*, the solid phase phosphoramidite triester method first described by Beaucage & Caruthers, *Tetrahedron Letts.* 22:1859-1862 (1981),  
30 using an automated synthesizer, as described in Van Devanter *et. al.*, *Nucleic Acids Res.* 12:6159-6168 (1984). Purification of oligonucleotides is performed using known methods,

*e.g.*, by native acrylamide gel electrophoresis or by anion-exchange HPLC as described in Pearson & Reanier, *J. Chrom.* 255:137-149 (1983).

[0031] The nucleic acids encoding the hybrid proteins or segments of the hybrid proteins can be amplified from nucleic acid samples, *e.g.*, oligonucleotide segments, using various  
5 amplification/extension techniques. For instance, polymerase chain reaction (PCR) technology may be used to obtain nucleic acid sequences that code for hybrid proteins to be expressed, to make nucleic acids to use as probes for detecting the presence of the desired nucleic acid sequences in samples, for nucleic acid sequencing, or for other purposes. For a general overview of PCR see PCR Protocols: A Guide to Methods and Applications. (Innis,  
10 M, Gelfand, D., Sninsky, J. and White, T., eds.), Academic Press, San Diego (1990).

[0032] Nucleic acid sequences encoding the hybrid proteins of the invention can be cloned into expression vectors to generate a library of sequences encoding individual hybrid proteins.

[0033] The following discussion provides details on how to select and align the parent  
15 starting proteins, how to create the library of nucleic acids derived from the parent proteins, and how to screen the library for functional proteins.

## **B. Definitions**

[0034] The term "hybrid protein" is used herein to describe a protein that comprises amino  
20 acid residues from multiple parent sequences.

[0035] The term "amplification" refers to a process whereby the number of copies of a nucleic acid fragment is increased.

[0036] A "parent sequence" indicates a starting or reference amino acid or nucleic acid  
25 sequence prior to a manipulation of the invention. The term is used interchangeably with "starting sequence". Parent sequences may be wild-type proteins, hybrid proteins, proteins containing mutations, or other engineered proteins. Parent sequences can be full-length proteins, protein subunits, protein domains, amino acid motifs, protein active sites, or any polypeptide sequence or subset of polypeptide sequences, whether continuous or interrupted by other polypeptide sequences.

[0037] The term "wild-type" refers to a polynucleotide or polypeptide sequence that does not comprise mutations. A "wild-type" protein refers to a protein active at a level of activity found in nature and that typically comprises the amino acid sequence found in nature.

[0038] The term "mutations" refers to changes in the sequence of a wild-type nucleic acid sequence or changes in the sequences of a peptide. Such mutations may be point mutations such as transitions or transversions, or deletions, insertions, or duplications.

[0039] The term "naturally occurring" as used herein refers to a nucleic acid or polypeptide that can be found in nature. For example, a polypeptide or polynucleotide sequence that is present in an organism that can be isolated from a source in nature and which has not been intentionally modified in a laboratory is naturally-occurring.

[0040] A "common biological" activity refers to an activity that is shared by two or more proteins wherein the common biological activity is an activity that is found in nature. Biological activity of a protein can be assessed using standard means known in the art for determining the function of a protein.

[0041] A polynucleotide sequence is "heterologous to" an organism or a second polynucleotide sequence if it originates from a foreign species, or, if from the same species, is modified from its original form. For example, a promoter operably linked to a heterologous coding sequence refers to a coding sequence from a species different from that from which the promoter was derived, or, if from the same species, a coding sequence which is different from any naturally occurring allelic variants.

[0042] An "expression cassette" refers to a polynucleotide with a series of nucleic acid elements that permit transcription and often translation of a particular nucleic acid, *e.g.* in a cell. Typically, the expression cassette includes a nucleic acid to be transcribed operably linked to a promoter and a ribosomal binding site.

[0043] "Nucleic acid" and "polynucleotide" are used interchangeably herein to refer to deoxyribonucleotides or ribonucleotides and polymers thereof in either single- or double-stranded form. The term encompasses nucleic acids containing known nucleotide analogs or modified backbone residues or linkages, which are synthetic, naturally occurring, and non-naturally occurring, which have similar binding properties as the reference nucleic acid, and which are metabolized in a manner similar to the reference nucleotides. Examples of such analogs include, without limitation, phosphorothioates, phosphoramidates, methyl

phosphonates, chiral-methyl phosphonates, 2-O-methyl ribonucleotides, and peptide-nucleic acids (PNAs).

[0044] "Polypeptide," "peptide" and "protein" are used interchangeably herein to refer to a polymer of amino acid residues. The terms apply to naturally occurring amino acid  
5 polymers, as well as amino acid polymers in which one or more amino acid residue is an artificial chemical mimetic of a corresponding naturally occurring amino acid.

[0045] Two nucleic acid sequences or polypeptides are said to be "identical" if the sequence of nucleotides or amino acid residues, respectively, in the two sequences is the same when aligned for maximum correspondence as described below. The terms "identical"  
10 or percent "identity," in the context of two or more nucleic acids or polypeptide sequences, refer to two or more sequences or subsequences that are the same or have a specified percentage of amino acid residues or nucleotides that are the same, when compared and aligned for maximum correspondence over a comparison window, as measured using one of the sequence comparison algorithms described later or by manual alignment and visual  
15 inspection.

[0046] When referring to proteins or peptides and for purposes of aligning polypeptides, it is recognized that residue positions that are not identical often differ by conservative amino acid substitutions, where amino acids residues are substituted for other amino acid residues with similar chemical properties (*e.g.* charge or hydrophobicity) and do not necessarily  
20 change the functional properties of the molecule. The scoring of conservative substitutions for the purposes of this patent is based on the BLOSUM62 matrix (Henikoff & Henikoff, *Proc. Natl. Acad. Sci. USA* 89:10915, 1989, Figure 1).

[0047] The term "sequence similarity" or "similar" may also be used with respect to amino acid sequences. This term encompasses conservative substitutions, as described above. For  
25 purposes of determining percent similarity, two amino acids are considered similar if they are given a value greater than zero (0) on the BLOSUM62 substitution matrix (Figure 1). Optimal alignment for determining percent sequence similarity may be conducted using various algorithms as further explained hereinbelow. In cases where an optimal alignment of two sequences requires the insertion of a gap in one or both of the sequences, an amino acid  
30 residue in one sequence that aligns with a gap in the other sequence is counted as a mismatch for purposes of determining percent identity. Gaps can be internal or external, *i.e.*, a truncation.

[0048] The term "absolute percent identity" refers to a percentage of sequence identity determined by scoring identical amino acids as 1 and any substitution as zero, regardless of the similarity of mismatched amino acids. In a typical sequence alignment, *e.g.* a BLAST alignment, the "absolute percent identity" of two sequences is presented as a percentage of amino acid "identities." As used herein, where a sequence is defined as being "at least X% identical" to a reference sequence, *e.g.* "a polypeptide at least 90% identical to SEQ ID NO:2," it is to be understood that "X% identical" refers to absolute percent identity, unless otherwise indicated. In cases where an optimal alignment of two sequences requires the insertion of a gap in one or both of the sequences, an amino acid residue in one sequence that aligns with a gap in the other sequence is counted as a mismatch for purposes of determining percent identity. Gaps can be internal or external, *i.e.*, a truncation.

[0049] The term "substantial identity" or "substantial similarity" of polynucleotide sequences means that a polynucleotide comprises a sequence that has at least 25% sequence identity, or sequence similarity, respectively. Alternatively, percent identity or percent similarity can be any integer from at least 25% to 100% (*e.g.* at least 25%, 26%, 27%, 28%, 29%, 30%, 31%, 32%, 33%, 34%, 35%, 36%, 37%, 38%, 39%, 40%, 41%, 42%, 43%, 44%, 45%, 46%, 47%, 48%, 49%, 50%, 51%, 52%, 53%, 54%, 55%, 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 64%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, 100%). Some embodiments include at least: 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, or 99% compared to a reference sequence using the programs described herein; preferably BLAST using standard parameters, as described below. One of skill will recognize that these values can be appropriately adjusted to determine corresponding identity or similarity of proteins encoded by two nucleotide sequences by taking into account codon degeneracy, amino acid similarity, reading frame positioning and the like, as further described below.

[0050] Another indication that two nucleic acid sequences are substantially identical is that the two molecules or their complements hybridize to each other under stringent conditions, as described below. The phrase "selectively (or specifically) hybridizes to" refers to the binding, duplexing, or hybridizing of a molecule only to a particular nucleotide sequence under stringent hybridization conditions when that sequence is present in a complex mixture (*e.g.* total cellular or library DNA or RNA).

[0051] The phrase "stringent hybridization conditions" refers to conditions under which a probe will hybridize to its target subsequence, typically in a complex mixture of nucleic acid, but to no other sequences. Stringent conditions are sequence-dependent and will be different in different circumstances. Longer sequences hybridize specifically at higher temperatures.

5 An extensive guide to the hybridization of nucleic acids is found in Tijssen, Techniques in Biochemistry and Molecular Biology--Hybridization with Nucleic Probes, "Overview of principles of hybridization and the strategy of nucleic acid assays" (1993). Generally, highly stringent conditions are selected to be about 5-10°C lower than the thermal melting point (T<sub>m</sub>) for the specific sequence at a defined ionic strength pH. Low stringency conditions are  
10 generally selected to be about 15-30°C below the T<sub>m</sub>. The T<sub>m</sub> is the temperature (under defined ionic strength, pH, and nucleic concentration) at which 50% of the probes complementary to the target hybridize to the target sequence at equilibrium (as the target sequences are present in excess, at T<sub>m</sub>, 50% of the probes are occupied at equilibrium). Stringent conditions will be those in which the salt concentration is less than about 1.0 M  
15 sodium ion, typically about 0.01 to 1.0 M sodium ion concentration (or other salts) at pH 7.0 to 8.3 and the temperature is at least about 30°C for short probes (*e.g.* 10 to 50 nucleotides) and at least about 60°C for long probes (*e.g.* greater than 50 nucleotides). Stringent conditions may also be achieved with the addition of destabilizing agents such as formamide. For selective or specific hybridization, a positive signal is at least two times background,  
20 preferably 10 times background hybridization.

[0052] In the present invention, nucleic acids encoding polypeptides of the invention can be identified in standard Southern blots under stringent conditions using the nucleic acid sequences disclosed here. For the purposes of this disclosure, suitable stringent conditions for such hybridizations are those which include a hybridization in a buffer of 40%  
25 formamide, 1 M NaCl, 1% SDS at 37°C, and at least one wash in 0.2X SSC at a temperature of at least about 50°C, usually about 55°C to about 60°C or 60°C, for 20 minutes, or equivalent conditions. A positive hybridization is at least twice background. Those of ordinary skill will readily recognize that alternative hybridization and wash conditions can be utilized to provide conditions of similar stringency.

30 [0053] A further indication that two polynucleotides are substantially identical is if the reference sequence, amplified by a pair of oligonucleotide primers, can then be used as a probe under stringent hybridization conditions to isolate the test sequence from a cDNA or genomic library, or to identify the test sequence in, *e.g.* a northern or Southern blot.

[0054] The terms “library members,” “members of a the library” and the like refer to those nucleic acids present in a nucleic acid library that have the intended characteristics described herein; that is, nucleic acids that encode hybrid polypeptides. Small library members, *e.g.*, comprising sequences encoding polypeptide domains, can of course be joined with other library members or parental or non-parental sequences to encode full length proteins. It is recognized that libraries may in addition contain other nucleic acids, either as intentional additions or unintended contaminants; these additional nucleic acids are not considered “members”.

[0055] The terms ‘minimal encoding sequence’, ‘minimal encoding oligonucleotide sequence’, and ‘minimal encoding nucleotide sequence’ refer to nucleotide sequences that encode a library of hybrid sequences. It is the result of examining two or more different amino acid sequences and deducing a single degenerate nucleotide sequence that will encode a library of proteins that includes hybrid proteins derived from the two different amino acid sequences. Minimal encoding sequences can refer to a single codon, several codons or enough codons to encode an entire protein. Minimal encoding sequences need not be continuous. Minimal encoding sequences may encode non-parental amino acids both similar and dissimilar to parental sequences. Often, it is possible to deduce multiple minimal encoding sequences that can encode the same parental amino acids.

### C. Parent Protein Selection and alignment of sequences

[0056] In the methods of the invention, at least two polynucleotide sequences encoding polypeptides with a common biological activity (*e.g.*, deoxyuridine triphosphate nucleotidohydrolases, or DNA polymerases) are recombined to produce a library of hybrid polynucleotides. The library is then screened to identify functional hybrid proteins with an altered phenotype relative to the parent polypeptides.

[0057] The parent proteins can show substantial sequence or secondary structural similarity with each other, but they should also differ in at least 5 positions and may differ by as many 20, 50, 100, or 200, or more positions. The percent similarity or percent identity between the parent proteins can be any number from at least 60% to 99%. In comparing the initial sequences, there may be more than two parents. The multiple sequences may be divergent at a single position or at different positions. For example, there may be three related sequences that are the parents for generating hybrid molecules. One sequence may differ from the

second at a single position, and the second may differ from the third at a different single position.

[0058] The majority of differences, *e.g.*, greater than 50%, often greater than 75% or 90% of the differences, between the library members are typically confined to those corresponding amino acids that differ between the parent proteins. A corresponding amino acid refers to an amino acid residue of a parent sequence that occurs at a particular position when the parent sequences are maximally aligned. It should be understood that such position designations do not indicate the number of amino acids in the parent sequences *per se*, but indicate where in the parent sequence the residues occur. Alignment can be performed either manually or using a sequence comparison algorithm, as further explained below. For example, Figure 2 shows the amino acid sequences of two wild-type proteins that can be used in the methods of the present invention, Pfu and Deep Vent® polymerases. Figure 2 also indicates the positions of amino acids that differ between the two parent proteins. Typically, the hybrid proteins created by the methods of the invention will differ from each other in the positions that differ between their parents.

[0059] The initial differences in sequence between the parent proteins are typically, but not necessarily, the result of natural variation. For example, the parent proteins can be variant forms that are obtained from different individuals or strains of an organism, *e.g.*, the parent proteins can be related sequences from the same organism (*e.g.*, paralogs or allelic variations), or can be homologs from different organisms (interspecies homologs).

[0060] Accordingly, the parent polypeptides used in the methods of this invention are any set of two or more homologous proteins that share a common biological activity. Biological activity is not always shown directly, but may be inferred from sequence similarity or identity to known proteins of demonstrated activity. Biological activity may refer to a single enzyme activity even if a particular protein may have more than one enzymatic activity. Also, biological activity may refer to non-enzymatic activities such as binding to another molecule with allosteric effect, as with a hormone or a receptor; or binding to another molecule without allosteric effect, as for certain antibodies; or binding to another molecule with the effect of neutralization or sequestration.

[0061] Biological activity of a protein can be assessed using standard means known in the art for determining the function of a protein. For example, in some embodiments, the parent proteins will be enzymes and will share a common enzymatic activity. Exemplary enzymes

include polymerases, ligases, lipases, dehydrogenases, RNAses, DNAses, proteases, kinases, caspases, methylases, transcription factors, and restriction endonucleases. In other embodiments, the parent proteins will be other proteins, *e.g.*, receptors, hormones, immunoglobulins, or chromophores. The biological activity of these types of proteins can be assessed using known assays. The skilled practitioner will understand that any protein set, where the protein members are homologous proteins having at least 60% amino acid similarity, and often at least 60% identity, to each other and having a common biological activity, can be used as parental polypeptides.

[0062] The parent protein sequences are aligned according to standard methods. The sequences are compared and aligned for maximum correspondence over a comparison window, or designated region as measured using one of the sequence comparison algorithms described below or by manual alignment and visual inspection. The parent protein sequences can be aligned using any of the known algorithms suitable for determining percent sequence identity and sequence similarity. For purposes of this patent, percent amino acid identity and percent amino acid similarity is determined by the default parameters of BLASTP using the Blosum62 similarity matrix, an expectation of 10, a word size of 3, and a gap costs setting of existence 11/extension 1 (Altschul *et al.*, *Nuc. Acids Res.* 25:3389-3402 (1977)).

[0063] For sequence comparison, typically one sequence acts as a reference sequence, to which test sequences are compared. When using a sequence comparison algorithm, test and reference sequences are entered into a computer, subsequence coordinates are designated, if necessary, and sequence algorithm program parameters are designated. Default program parameters can be used, or alternative parameters can be designated. The sequence comparison algorithm then calculates the percent sequence identities or sequence similarities for the test sequences relative to the reference sequence, based on the program parameters.

[0064] The comparison window includes reference to a segment of any one of the number of contiguous positions selected from the group consisting of from 10 to 600, usually about 50 to about 200, more usually about 100 to about 150 in which a sequence may be compared to a reference sequence of the same number of contiguous positions after the two sequences are optimally aligned. Methods of alignment of sequences for comparison are well-known in the art. Optimal alignment of sequences for comparison can be conducted, *e.g.* by the local homology algorithm of Smith & Waterman, *Adv. Appl. Math.* 2:482 (1981), by the homology alignment algorithm of Needleman & Wunsch, *J. Mol. Biol.* 48:443 (1970), by the

search for similarity method of Pearson & Lipman, *Proc. Nat'l. Acad. Sci. USA* 85:2444 (1988), by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Dr., Madison, WI), or by manual alignment and visual inspection (see, *e.g.* Current Protocols in Molecular Biology (Ausubel *et al.*, eds. 1995 supplement)).

[0065] An example of an algorithm that is suitable for determining percent sequence identity and sequence similarity are the BLAST and BLAST 2.0 algorithms, which are described in Altschul *et al.*, *Nuc. Acids Res.* 25:3389-3402 (1977) and Altschul *et al.*, *J. Mol. Biol.* 215:403-410 (1990), respectively. Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov/>). This algorithm involves first identifying high scoring sequence pairs (HSPs) by identifying short words of length W in the query sequence, which either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a database sequence. T is referred to as the neighborhood word score threshold (Altschul *et al.*, *supra*). These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Cumulative scores are calculated using, for nucleotide sequences, the parameters M (reward score for a pair of matching residues; always > 0) and N (penalty score for mismatching residues; always < 0). For amino acid sequences, a scoring matrix is used to calculate the cumulative score. Extension of the word hits in each direction are halted when: the cumulative alignment score falls off by the quantity X from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached. The BLAST algorithm parameters W, T, and X determine the sensitivity and speed of the alignment. The BLASTN program (for nucleotide sequences) uses as defaults a wordlength (W) of 11, an expectation (E) of 10, M=5, N=-4 and a comparison of both strands. For amino acid sequences, the BLASTP program uses as defaults a wordlength of 3, and expectation (E) of 10, and the BLOSUM62 scoring matrix (see Henikoff & Henikoff, *Proc. Natl. Acad. Sci. USA* 89:10915 (1989)) alignments (B) of 50, expectation (E) of 10, M=5, N=-4, and a comparison of both strands.

[0066] The BLAST algorithm also performs a statistical analysis of the similarity between two sequences (see, *e.g.* Karlin & Altschul, *Proc. Nat'l. Acad. Sci. USA* 90:5873-5787

(1993)). One measure of similarity provided by the BLAST algorithm is the smallest sum probability ( $P(N)$ ), which provides an indication of the probability by which a match between two nucleotide or amino acid sequences would occur by chance. For example, a nucleic acid is considered similar to a reference sequence if the smallest sum probability in a comparison of the test nucleic acid to the reference nucleic acid is less than about 0.2, more preferably less than about 0.01, and most preferably less than about 0.001.

[0067] Proteins that are useful as parent proteins for generating protein hybrids typically have greater than 60% amino acid similarity, often greater than 60% identity, to each other. In some embodiments, the chosen parent sequences can be aligned without gaps. In other embodiments, the alignment must accommodate the presence of a gap or deletion in the amino acid sequence of one of the parent proteins.

[0068] Methods of aligning proteins containing gaps or deletions are known in the art. In some embodiments, the gap may be a result of a loop linking alpha helices or a turn in beta sheets. Typically, the gap or deletion will not affect the shared enzymatic activity between the two parent proteins. In some embodiments of the present invention, the starting sequences will be aligned in such a way as to include the gap or deletion. Standard programs for protein structure modeling can be used to help determine whether to omit or include an amino acid where a gap exists in an alignment, *e.g.* BLASTP.

[0069] For example, after inputting a protein sequence into BLASTP, a list of proteins is created with percent identities, similarities, and gaps indicated across the top of each pairwise comparison. If a gap is indicated, a library can be made to have some representatives with the gap and some without the gap.

[0070] In some embodiments, the gap may be of significant size, *e.g.* 2-50 amino acids, or have a loop with an enzymatic role. By aligning the two parent proteins to include the diversity of a gap or loop, larger diversity can be achieved.

[0071] Related proteins frequently have differences in length at their amino and carboxyl termini. Thus, the parental protein sequences may have length differences at their amino and carboxyl termini. The additional amino acids at either end may or may not contain a motif essential for function. For example, the carboxyl end of some type B polymerases contains a proliferating cell nuclear antigen (PCNA) binding motif. In some embodiments, one or more of the parent proteins will retain its C or N terminus motif. A library can be made to have some representatives with a C or N terminus tail and some without a C or N terminus tail.

[0072] Accordingly, the present invention provides methods of aligning parent sequences with or without gaps, deletions, or differences in their amino and carboxy terminals and combining them to construct a hybrid protein library and hybrid proteins of the present invention.

5

#### **D. Creating Polynucleotide Hybrids**

[0073] After the parent proteins have been selected and aligned, the mismatches between the sequences are identified. Hybrid oligonucleotide sequences are then generated that contain a mixture of parental residues at the mismatched sites, *i.e.*, for any given hybrid sequences, some of the residues at some of the mismatched sites are from one parent; residues at other of the mismatched sites are from another parent. A library comprising hybrid sequences can thereby be created. Considerations in generating the hybrid molecule libraries are set forth below.

10

##### *Codon selection*

[0074] After the amino acid sequences of the homologous parent proteins are aligned, the amino acids residues that are different between the sequences are identified. For each set of different amino acid residues, the codons that encoded the differing residues are compared and a minimal encoding sequence is derived. Preferably, codons that encode the different residues and only differ by one nucleotide are then selected as a point of degeneracy, *i.e.*, a point at which nucleotide variation results in codons encoding only one or the other parental amino acid.

15

20

[0075] Typically, the derivation of the minimal encoding sequence is also governed by the codon usage of a particular host. For example, if a nucleic acid encoding a hybrid protein is to be expressed in *E. coli*, the codon usage of *E. coli* can be used to derive a polynucleotide sequence that comprises preferred *E. coli* codons. An *E. coli* codon usage table can thus be used to compare the various codons that can encode two amino acids that differ.

25

[0076] In the simplest, and typically most common, case a single nucleic acid degeneracy can encode both amino acids that differ at a particular position in the parent sequences. For example, two homologous proteins may differ at a particular position where one parent has a valine at the position and the other has an isoleucine residue. Valine can be encoded by a number of different codons, one of which is GTT. Isoleucine can be encoded by a number of different residues, one of which is ATT. Therefore a minimal encoding sequence is (G/A)TT

30

(or RTT using the standard single letter code). The first nucleotide of the codon is the site of degeneracy. Oligonucleotide synthesis machines can readily be directed to produce product with half G and half A at a particular position. Individual nucleic acid molecules generated during the synthesis will therefore have either G or A at that particular site and the library of hybrid sequences will have some sequences with a G for this codon and some with an A. Accordingly, the proteins encoded by the individual library members will have either a valine or an isoleucine at that site. The degeneracy created at this site is independent of degeneracies created at other sites. This results in a library with a large number of variants, but that is constrained by the sequences of the parental polypeptides.

**[0077]** In comparing some of the differences where a position in the aligned sequences has two different amino acids, a minimal encoding sequence may require two nucleotides to be changed in order to encode the two parental residues at that position. This can result in a situation where two non-parental amino acid sequences may also be encoded by the degenerate codon. For example, two parental sequences may differ at a particular position, where one residue is a lysine and the other parental residue is an alanine. Lysine is encoded by AAR and Ala by GCN. The minimal encoding sequence (A/G)(A/C)G may therefore be used to encode both lysine and alanine. However, such a degenerate codon also can encode threonine (ACN) and glutamine (GAR) in addition to Lys and Ala. In some instances, a hybrid protein may tolerate an amino acid residue that is not in either parent, especially if the non-parental amino acid is similar to the one of the parental amino acids. In other instances, for example, if the sequence occurs in a domain that is known to be important for protein activity, it may not be desirable to introduce non-parental amino acid residues. Moreover, in some cases, the degenerate codon could result in the introduction of a stop codon, which would result in a library in which a portion of the sequences are not useful. Typically, one of the parent residues is selected for this position.

**[0078]** The purpose of producing the hybrid library should be considered in making the decision as to which parent residue to choose. For instance, if a desired function such as thermal stability, or exonuclease activity level, is greater in one parent than another, the choice should favor the parent with the desired characteristic. All things being equal, in cases where there are more than two parental protein sequences or there are additional isozymes, homologues, or related sequences, decisions on whether to include a particular amino acid may be made by “voting” – for example, if fewer than a threshold fraction of the parental

sequences differ in one position from an amino acids present in a majority of the sequences, then the rare amino acid may be ignored.

[0079] This situation may also be addressed by generating two different nucleic acid sequences, *e.g.*, by synthesizing two different oligonucleotides, one of which encodes one of the parent residues, the other of which encodes the different parent residue. For the purposes of generating a library, mixing the two oligonucleotides in equal amounts will effectively produce a degenerate oligonucleotide encoding the two amino acids exclusively. This mixed nucleic acid sequence may then be used for assembly of nucleic acids encoding a hybrid protein library.

[0080] In some cases, the minimal encoding sequence is more likely to encode non-parental amino acid sequences because they share no codon sequences in common (*e.g.*, Met and Asp). In this case again, the non-parental amino acids can be accepted as mutations in the library, or one of the parent codon sequence can be selected for incorporation into the hybrid protein library at this position, or two libraries can be constructed and combined as described above, or, if more than two parental sequences are used, or homologues are known, then the decision may be made by “voting”.

[0081] In comparing two homologous sequences, gaps and deletions may also occur. As the parental proteins share a common activity, the gaps typically do not significantly affect activity. For example, the homologous proteins may include loops linking alpha helices or turns in beta sheets. The absolute size of these links and turns are often not consequential. In general, a gap may be accommodated in generating the hybrid library by generating two versions of the nucleic acid sequence, *e.g.* synthesizing two oligonucleotides, mixing the two sequences, and using the mixture to construct the library. Alternatively, the hybrid library can be constructed so that the gap is present in all of the members or absent in all of the members. Similarly, related proteins frequently have differences in length at their amino and carboxyl termini. Again, two sequence may be generated, one with a longer end and one without the longer end, and then combined; or a library can be generated that includes the additional length or omits the additional length; or, if more than two parental sequences are used, then the decision may be made by “voting”.

[0082] Often, it may be desirable to introduce restriction endonuclease sites into the sequences of the library, for example, in order to facilitate assembly of the sequence encoding the protein, or domain exchange. One of skill in the art understands that such sites are

typically relatively infrequent, *e.g.*, have a 6 base pair recognition site. The restriction sites are often introduced into the nucleic acids by modifying codons without changing the amino acid encoded by the codon. The restriction sites are typically introduced into regions of the two parent sequences that are identical, although this need not be the case.

5 *Preparation of hybrid sequences and library production*

[0083] After the minimal encoding sequence is selected, the library is constructed using techniques well known in the art. Typically, the nucleic acids to be incorporated into the library are synthesized as oligonucleotides that are assembled to form a sequence encoding the hybrid polypeptide. Procedures for performing this are well known in the art.

10 Oligonucleotides of about 50-100 bases are typically synthesized. The oligonucleotides are designed such that they overlap, *e.g.*, by 10 to 50 bases, to provide adequate annealing and specificity despite the sequence differences. As appreciated by one of skill, the 3' ends often are in regions in which there are minimal or no differences between the parent sequences.

[0084] The completed gene is then assembled, *e.g.*, by primer extension (*see, e.g.*, Figure 15 3). In such an assembling procedure overlapping oligonucleotides are annealed to one another and extended using a high fidelity thermostable polymerase. Large amounts of primer and minimal cycles (usually between 0 and 5) are used in assembling segments. The products are then purified and used for the next cycle of pairing and primer extension.

[0085] The resulting reassembled polynucleotide can be of various lengths. Preferably the 20 reassembled sequences are from about 50 bp to about 10 kb.

[0086] As appreciated by one of skill, the gene encoding the hybrid polypeptide can also be assembled by ligating the appropriate fragments. Further, the full-length hybrid polypeptide can be assembled by ligating together the appropriate smaller fragments. If the hybrid polypeptide is a portion of a larger protein, incorporation into the larger protein can also 25 occur at this step. Often, restriction endonuclease sites can be incorporated into the primers to improve the efficiency of the ligation step.

[0087] In some instances, it is desirable to prepare two libraries and then combine them, for example, in instances in which there is a gap in the parental sequences or two amino acid residues that differ in the parent sequences differ in their codons at all three nucleotide 30 position.

[0088] As appreciated by those in the art, the hybrid molecules can further be used as substrates to generate additional diversity by using various techniques such as recursive recombination (*see, e.g.*, U.S. Patent No. 6,180,406, and related patents); and various other mutagenesis procedures, *e.g.*, error-prone PCR, cassette mutagenesis. These techniques may be performed on all of the library members or selected subpopulation or individual library members.

[0089] In some recombination techniques, polynucleotide fragments are recombined by linking overlapping single stranded segments and then contacting the resulting linked segments with a polymerase. *See, e.g.* U. S. Patent No. 6,150,111.

[0090] In other techniques, recombination is independent of natural restriction sites or *in vitro* ligation (Ma *et al.*, *Gene* 58:201-216 (1989); Oldenburg *et al.*, *Nucleic Acids Research* 25:451-452 (1997)). In some of these methods, an *in vivo* method for plasmid construction takes advantage of the double-stranded break repair pathway in a cell such as a yeast cell to achieve precision joining of DNA fragments. This method involves synthesis of linkers, *e.g.* 60-140 base pairs, from short oligonucleotides and requires assembly by enzymatic methods into the linkers needed (Raymond *et al.*, *BioTechniques* 26(1):134-141 (1999)).

[0091] In some techniques, short random or non-random oligonucleotide sequences are recombined with polynucleotide segments derived from polynucleotides encoding functional polymerases.

[0092] Modifications may also be introduced into the polynucleotide segments or the assembled polynucleotides encoding the hybrid proteins using other known mutagenesis techniques. For example, the polynucleotides can be submitted to one or more rounds of error-prone PCR (*e.g.* Leung, D. W. *et al.*, *Technique* 1:11-15 (1989); Caldwell, R. C. and Joyce, G. F. *PCR Methods and Applications* 2:28-33 (1992); Gramm, H. *et al.*, *Proc. Natl. Acad. Sci. USA* 89:3576-3580 (1992)), thereby introducing variation into the polynucleotides. Alternatively, cassette mutagenesis (*e.g.* Stemmer, W. P. C. *et al.*, *Biotechniques* 14:256-265 (1992); Arkin, A. and Youvan, D. C. *Proc. Natl. Acad. Sci. USA* 89:7811-7815 (1992); Oliphant, A. R. *et al.*, *Gene* 44:177-183 (1986); Hermes, J. D. *et al.*, *Proc. Natl. Acad. Sci. USA* 87:696-700 (1990)), in which the specific region to be optimized is replaced with a synthetically mutagenized oligonucleotide, can be used. Mutator strains of host cells can also be employed to add to mutational frequency (Greener and Callahan, *Strategies in Mol. Biol.* 7: 32 (1995)).

[0093] Site directed mutagenesis is well known in the art and may also be used to introduce further diversity into the sequences. Such techniques include site directed mutagenesis as described, *e.g.*, in Ling *et al.* (1997) *Anal Biochem.* 254(2): 157-178; Dale *et al.* (1996) *Methods Mol. Biol.* 57:369-374; Smith (1985) *Ann. Rev. Genet.* 19:423-462; Botstein & Shortle (1985) *Science* 229:1193-1201; Carter (1986) *Biochem. J.* 237:1-7; and Kunkel (1987) "The efficiency of oligonucleotide directed mutagenesis" in *Nucleic Acids & Molecular Biology* (Eckstein, F. and Lilley, D.M.J. eds., Springer Verlag, Berlin)); mutagenesis using uracil containing templates (Kunkel (1985) *Proc. Natl. Acad. Sci. USA* 82:488-492; Kunkel *et al.* (1987) *Methods in Enzymol.* 154, 367-382; and Bass *et al.* (1988) *Science* 242:240-245); oligonucleotide-directed mutagenesis (*Methods in Enzymol.* 100: 468-500 (1983); *Methods in Enzymol.* 154: 329-350 (1987); Zoller & Smith (1982) *Nucleic Acids Res.* 10:6487-6500; Zoller & Smith (1983) *Methods in Enzymol.* 100:468-500; and Zoller & Smith (1987) *Methods in Enzymol.* 154:329-350); phosphorothioate-modified DNA mutagenesis (Taylor *et al.* (1985) *Nucl. Acids Res.* 13: 8749-8764; Taylor *et al.* (1985) *Nucl. Acids Res.* 13: 8765-8787 (1985); Nakamaye & Eckstein (1986) *Nucl. Acids Res.* 14: 9679-9698; Sayers *et al.* (1988) *Nucl. Acids Res.* 16:791-802; and Sayers *et al.* (1988) *Nucl. Acids Res.* 16: 803-814); mutagenesis using gapped duplex DNA (Kramer *et al.* (1984) *Nucl. Acids Res.* 12: 9441-9456; Kramer & Fritz (1987) *Methods in Enzymol.* 154:350-367; Kramer *et al.* (1988) *Nucl. Acids Res.* 16: 7207; and Fritz *et al.* (1988) *Nucl. Acids Res.* 16: 6987-6999).

[0094] An additional modification method well known in the art is point mismatch repair, *e.g.* (Kramer *et al.* (1984) *Cell* 38:879-887), mutagenesis using repair-deficient host strains (Carter *et al.* (1985) *Nucl. Acids Res.* 13: 4431-4443; and Carter (1987) *Methods in Enzymol.* 154: 382-403), deletion mutagenesis (Eghtedarzadeh & Henikoff (1986) *Nucl. Acids Res.* 14: 5115), restriction-selection and restriction-selection and restriction-purification (Wells *et al.* (1986) *Phil. Trans. R. Soc. Lond. A* 317: 415-423), mutagenesis by total gene synthesis (Nambiar *et al.* (1984) *Science* 223: 1299-1301; Sakamar and Khorana (1988) *Nucl. Acids Res.* 14: 6361-6372; Wells *et al.* (1985) *Gene* 34:315-323; and Grundström *et al.* (1985) *Nucl. Acids Res.* 13: 3305-3316), double-strand break repair (Mandecki (1986); Arnold (1993) *Current Opinion in Biotechnology* 4:450-455; *Proc. Natl. Acad. Sci. USA*, 83:7177-7181). Additional details on many of the above methods can be found in *Methods in Enzymology* Volume 154, which also describes useful controls for trouble-shooting problems with various mutagenesis methods.

[0095] The assembled gene fragments may then be cloned into any of a number of vectors to generate a library comprising individual hybrid molecules that comprise residues from the parent sequences.

*Expression of hybrid protein libraries*

5 [0096] There are many expression systems for producing the hybrid polypeptides and polypeptide libraries that are well known to those of ordinary skill in the art. (See, e.g., Gene Expression Systems, Fernandez and Hoeffler, Eds. Academic Press, 1999; Sambrook & Russell, *supra*; and Ausubel *et al, supra*.) Typically, the polynucleotide that encodes a hybrid polypeptide is placed under the control of a promoter that is functional in the desired  
10 host cell. An extremely wide variety of promoters are available, and can be used in the expression vectors of the invention, depending on the particular application. Ordinarily, the promoter selected depends upon the cell in which the promoter is to be active. Other expression control sequences such as ribosome binding sites, transcription termination sites and the like are also optionally included. Constructs that include one or more of these control  
15 sequences are termed "expression cassettes." Accordingly, the nucleic acids that encode the joined polypeptides are incorporated for high level expression in a desired host cell.

[0097] Expression control sequences that are suitable for use in a particular host cell are often obtained by cloning a gene that is expressed in that cell. Commonly used prokaryotic control sequences, which are defined herein to include promoters for transcription initiation,  
20 optionally with an operator, along with ribosome binding site sequences, include such commonly used promoters as the beta-lactamase (penicillinase) and lactose (*lac*) promoter systems (Change *et al.*, *Nature* (1977) 198: 1056), the tryptophan (*trp*) promoter system (Goeddel *et al.*, *Nucleic Acids Res.* (1980) 8: 4057), the *tac* promoter (DeBoer, *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* (1983) 80:21-25); and the lambda-derived P<sub>L</sub> promoter and N-gene  
25 ribosome binding site (Shimatake *et al.*, *Nature* (1981) 292: 128). The particular promoter system is not critical to the invention, any available promoter that functions in prokaryotes can be used. Standard bacterial expression vectors include plasmids such as pBR322-based plasmids, e.g., pBLUESCRIPT™, pSKF, pET23D, λ-phage derived vectors, and fusion  
30 expression systems such as GST and LacZ. Epitope tags can also be added to recombinant proteins to provide convenient methods of isolation, e.g., c-myc, HA-tag, 6-His tag, maltose binding protein, VSV-G tag, anti-DYKDDDDK tag, or any such tag, a large number of which are well known to those of skill in the art.

[0098] For expression of hybrid polypeptides in prokaryotic cells other than *E. coli*, a promoter that functions in the particular prokaryotic species is required. Such promoters can be obtained from genes that have been cloned from the species, or heterologous promoters can be used. For example, the hybrid *trp-lac* promoter functions in *Bacillus* in addition to *E.*  
5 *coli*. These and other suitable bacterial promoters are well known in the art and are described, *e.g.*, in Sambrook *et al.* and Ausubel *et al.* Bacterial expression systems for expressing the proteins of the invention are available in, *e.g.*, *E. coli*, *Bacillus sp.*, and *Salmonella* (Palva *et al.*, *Gene* 22:229-235 (1983); Mosbach *et al.*, *Nature* 302:543-545 (1983). Kits for such expression systems are commercially available.

10 [0099] Eukaryotic expression systems for mammalian cells, yeast, and insect cells are well known in the art and are also commercially available. In yeast, vectors include Yeast Integrating plasmids (*e.g.*, YIp5) and Yeast Replicating plasmids (the YRp series plasmids) and pGPD-2. Expression vectors containing regulatory elements from eukaryotic viruses are typically used in eukaryotic expression vectors, *e.g.*, SV40 vectors, papilloma virus vectors,  
15 and vectors derived from Epstein-Barr virus. Other exemplary eukaryotic vectors include pMSG, pAV009/A+, pMTO10/A+, pMAMneo-5, baculovirus pDSVE, and any other vector allowing expression of proteins under the direction of the CMV promoter, SV40 early promoter, SV40 later promoter, metallothionein promoter, murine mammary tumor virus promoter, Rous sarcoma virus promoter, polyhedrin promoter, or other promoters shown  
20 effective for expression in eukaryotic cells.

[0100] Either constitutive or regulated promoters can be used in the present invention. Regulated promoters can be advantageous because the host cells can be grown to high densities before expression of the polypeptides is induced. Further, high level expression of heterologous proteins may slow cell growth in some situations. An inducible promoter is a  
25 promoter that directs expression of a gene where the level of expression is alterable by environmental or developmental factors such as, for example, temperature, pH, anaerobic or aerobic conditions, light, transcription factors and chemicals.

[0101] For *E. coli* and other bacterial host cells, inducible promoters are known to those of skill in the art. These include, for example, the *lac* promoter, the bacteriophage lambda P<sub>L</sub>  
30 promoter, the hybrid *trp-lac* promoter (Amann *et al.* (1983) *Gene* 25: 167; de Boer *et al.* (1983) *Proc. Nat'l. Acad. Sci. USA* 80: 21), and the bacteriophage T7 promoter (Studier *et*

*al.* (1986) *J. Mol. Biol.*; Tabor *et al.* (1985) *Proc. Nat'l. Acad. Sci. USA* 82: 1074-8). These promoters and their use are discussed in Sambrook *et al.*, *supra*.

[0102] Inducible promoters for other organisms are also well known to those of skill in the art. These include, for example, the metallothionein promoter, the heat shock promoter, as well as many others.

[0103] Translational coupling may also be used to enhance expression. The strategy uses a short upstream open reading frame derived from a highly expressed gene native to the translational system, which is placed downstream of the promoter, and a ribosome binding site followed after a few amino acid codons by a termination codon. Just prior to the termination codon is a second ribosome binding site, and following the termination codon is a start codon for the initiation of translation. The system dissolves secondary structure in the RNA, allowing for the efficient initiation of translation. See Squires, *et. al.* (1988), *J. Biol. Chem.* 263: 16297-16302.

[0104] The construction of polynucleotide constructs generally requires the use of vectors able to replicate in bacteria. Such vectors are commonly used in the art. A plethora of kits are commercially available for the purification of plasmids from bacteria (for example, EasyPrepJ, FlexiPrepJ, from Pharmacia Biotech; StrataCleanJ, from Stratagene; and, QIAexpress Expression System, Qiagen). The isolated and purified plasmids can then be further manipulated to produce other plasmids, and used to transform cells.

[0105] The hybrid polypeptides can be expressed intracellularly, or can be secreted from the cell. Intracellular expression often results in high yields. If necessary, the amount of soluble, active polypeptide may be increased by performing refolding procedures (*see, e.g.*, Sambrook *et al.*, *supra.*; Marston *et al.*, *Bio/Technology* (1984) 2: 800; Schoner *et al.*, *Bio/Technology* (1985) 3: 151). Fusion polypeptides of the invention can be expressed in a variety of host cells, including *E. coli*, other bacterial hosts, yeast, and various higher eukaryotic cells such as the COS, CHO and HeLa cells lines and myeloma cell lines. The host cells can be mammalian cells, insect cells, or microorganisms, such as, for example, yeast cells, bacterial cells, or fungal cells.

[0106] Once expressed, the hybrid polypeptides can be purified according to standard procedures of the art, including ammonium sulfate precipitation, affinity columns, column chromatography, gel electrophoresis and the like (*see, generally*, R. Scopes, *Protein Purification*, Springer-Verlag, N.Y. (1982), Deutscher, *Methods in Enzymology Vol. 182*:

*Guide to Protein Purification.*, Academic Press, Inc. N.Y. (1990)). Substantially pure compositions of at least about 90 to 95% homogeneity are preferred, and 98 to 99% or more homogeneity are most preferred. Once purified, partially or to homogeneity as desired, the polypeptides may then be used (*e.g.*, as immunogens for antibody production).

5 [0107] To facilitate purification of the hybrid polypeptides of the invention, the nucleic acids that encode the fusion polypeptides can also include a coding sequence for an epitope or "tag" for which an affinity binding reagent is available. Examples of suitable epitopes include the myc and V-5 reporter genes; expression vectors useful for recombinant production of fusion polypeptides having these epitopes are commercially available (*e.g.*,  
10 Invitrogen (Carlsbad CA) vectors pcDNA3.1/Myc-His and pcDNA3.1/V5-His are suitable for expression in mammalian cells). Additional expression vectors suitable for attaching a tag to the fusion proteins of the invention, and corresponding detection systems are known to those of skill in the art, and several are commercially available (*e.g.*, FLAG" (Kodak, Rochester NY). Another example of a suitable tag is a polyhistidine sequence, which is  
15 capable of binding to metal chelate affinity ligands. Typically, six adjacent histidines are used, although one can use more or less than six. Suitable metal chelate affinity ligands that can serve as the binding moiety for a polyhistidine tag include nitrilo-tri-acetic acid (NTA) (Hochuli, E. (1990) "Purification of recombinant proteins with metal chelating adsorbents" In Genetic Engineering: Principles and Methods, J.K. Setlow, Ed., Plenum Press, NY;  
20 commercially available from Qiagen (Santa Clarita, CA)).

#### **E. Screening the hybrid library**

[0108] After the nucleic acid library is created using the methods described above, the library is screened for functional hybrids and/or hybrids that possess improved activity over their parents. A variety of assays known in the art that can be used to compare the activity of  
25 a hybrid protein to its wild-type counterpart.

[0109] The nature of screening or selection depends on the property or characteristic that is to be improved or acquired. Examples are provided below. It is not usually necessary to understand the molecular basis by which particular products of recombination (recombinant segments) have acquired new or improved properties or characteristics relative to the starting  
30 substrates.

[0110] Depending on the particular screening protocol used for a desired property, initial round(s) of screening can sometimes be performed using bacterial cells due to high

transfection efficiencies and ease of culture. However, for eukaryotic proteins, bacterial expression is often not practical, and yeast, fungal or other eukaryotic systems are used for library expression and screening. Similarly, other types of screening which are not amenable to screening in bacterial or simple eukaryotic library cells, are performed in cells selected for use in an environment close to that of their intended use. Final rounds of screening can be performed in the precise cell type of intended use.

[0111] If further diversity is desired, at least one, and usually a collection, of hybrid sequences that are identified in an initial screen/selection may be subjected to an additional round of hybrid generation or an additional procedure to generate diversity. For example, in generating the initial library, not all of the residues that are identified as differing in parental sequences may be the subject of hybrid generation, *i.e.*, particular amino acids present in a subset of the parental sequences may be selected. Subsequent rounds may be directed to generating hybrids comprising such additional residues. Further, an additional round of hybrid generation may be performed using a different parent sequence, *i.e.*, a parent sequence that was not included in the initial alignment. Lastly, a different diversity-generating procedure, *e.g.*, recursive recombination, may be used.

[0112] The second round of diversity generation can be followed by a further round of screening/selection according to the principles discussed above for the first round. The stringency of screening/selection may be increased between rounds. Also, the nature of the screen and the property being screened for may vary between rounds if improvement in more than one property is desired or if acquiring more than one new property is desired. Additional rounds of hybrid/diversity generation and screening can then be performed until the recombinant segments have sufficiently evolved to acquire the desired new or improved property or function.

[0113] For example, in some embodiments the parental sequences are polymerases and the hybrid proteins are selected for improved polymerase function, *e.g.*, processivity or error-correcting activity. These properties can be measured and compared to the parent polymerase activities using methodology well known in the art (*see, e.g.*, WO0192501).

[0114] Polymerase processivity is generally defined as the number of nucleotides incorporated during a single binding event of a modifying enzyme to a primed template. This activity may be assessed, *e.g.*, by using a procedure in which a 5' FAM-labeled primer is annealed to circular or linearized ssM13mp18 DNA to form a primed template. In measuring

processivity, the primed template usually is present in significant molar excess to the enzyme or catalytic domain to be assayed so that the chance of any primed template being extended more than once by the polymerase is minimized. The primed template is therefore mixed with the polymerase catalytic domain to be assayed at a ratio such as approximately 4000:1 (primed DNA:DNA polymerase) in the presence of buffer and dNTPs. MgCl<sub>2</sub> is added to initiate DNA synthesis. Samples are quenched at various times after initiation, and analyzed on a sequencing gel. At a polymerase concentration where the median product length does not change with time or polymerase concentration, the length corresponds to the processivity of the enzyme. The processivity of a hybrid protein of the invention is then compared to the processivity of one or more of the parent proteins.

[0115] Enhanced efficiency, *e.g.*, of a polymerase, can also be demonstrated by measuring the increased ability of an enzyme to produce product in fewer cycles. During a PCR reaction, samples can be removed from the reaction after each cycle and run on a gel to determine the quantity of product generated. More efficient enzymes will produce more product in the same number of cycles in comparison to a starting enzyme.

[0116] Of course, hybrid proteins, *e.g.*, receptor molecules, may be tested for an improved or acquired activity such as signaling or ligand binding using assays appropriate to the protein.

## EXAMPLES

[0117] These examples describe the generation of hybrid libraries and the isolation of hybrid proteins from the libraries.

### Example 1. Generation of hybrid Dut proteins

[0118] In this example the invention is to be used to isolate hybrid proteins with varying temperature optima. The model proteins are a mesophilic and thermophilic deoxyuridine 5'-triphosphate nucleotidohydrolase (dUTPase or Dut). Sequences of the mesophile *E. coli* Dut (ECD) and the thermophile *Aquifex aeolicus* Dut (AAD) genes were aligned using BlastP. The sequences are 40% identical and 60% similar as defined by the Blast default parameters. The sequences and Blast alignment are shown in Figure 4.

[0119] The aligned parental sequences and all possible codons in order of frequency of use by *E. coli* are shown in Figure 5A. A minimal encoding sequence was derived by identifying codons that will encode both sequences with a minimal number of degeneracies (Figure 5B). Codons frequently used by *E. coli* are preferred. There are 90 differences between the two

sequences. Of these, 49 can be encoded by incorporating a single degeneracy in the DNA sequence. Most of the others, 38 of them, require two degeneracies and 1 requires three. There is one gap. Two of the degeneracies could result in termination codons being incorporated into the sequence. Nucleic acid degeneracies that might incorporate termination sites or amino acids not similar (BLOSUM 62 number is  $<0$ ) to either parental amino acid sequence were removed (Figure 5C) and replaced with sequences of the more thermal stable AAD, in keeping with the experimental purpose. If this was not done, up to 24% of the amino acids incorporated into the hybrid protein may have been non-parental; some of those with no similarity to either parent. The elimination of the non-similar sequences reduced the maximum number of non-parental amino acids to 14%, all of which would be similar to at least one parent.

**[0120]** The double-stranded nucleic acid sequence showing the degeneracies and encoded amino acid residues is shown in Figure 6. Priming and restriction sites were added to the ends (shown in bold). In two cases, codon usage was changed to add restriction sites (underlined and in italics). The amino acids encoded by the sequence are indicated below the codons.

**[0121]** Figure 7 shows the full length hybrid library nucleic acid sequence. The degenerate positions are represented using standard single letter code. Oligonucleotides sequences were selected for synthesis (shown in bold). Selections were made such that minimal degeneracies exist where primers were expected to anneal to one another during assembly. In one portion of the sequence, there were no regions where reasonably-sized (about 10 to 50 bases), annealable oligonucleotides sequences could be selected. In this example, the *Cla*I site (underlined) inserted in the previous step is used to assemble a full length protein encoding library.

**[0122]** A hybrid library encoding a small protein such as DUT can be constructed by synthesizing the oligonucleotides so that there are no gaps once the primers are annealed. In this case, ligation rather than assembly PCR may be used to construct the hybrid library. The oligonucleotides are simply sequentially annealed, ligated, purified, than annealed again.

**[0123]** The final primers chosen in this example are indicated below. Assembly would occur as follows: Fwd1 primer is annealed to RevA primer. In separate tubes, Fwd2 is annealed to RevB, Fwd3 to RevC, Fwd4 to RevD, and Fwd5 is annealed to Rev5. The products of the five annealing reactions are primer extended with a DNA dependent DNA

polymerase with proof reading activity, typically *E. coli* DNA polymerase I Klenow fragment, or the thermal stable Phusion polymerase (MJ Research, Inc.). If Phusion polymerase is used, it is possible to thermally cycle the primer extension reaction. The products of the Fwd1/RevA reaction are annealed to the products of the Fwd2/RevB reaction and extension is repeated. Similarly, the products of the Fwd4/RevD reaction are annealed to the Fwd5/RevE reaction and extended. Finally, the Fwd1/RevA/Fwd2/RevB products are annealed to the Fwd3/RevC products and extended.

Oligonucleotides:

Fwd 1: 5'-TTGGTACCAA GCTTCATATG A-3'

10

Fwd 2: 5'-CC GCTGCCGASC TATGCGACCY CTCACAGCKC AGGCCTGGAT  
CTGCGTGCG-3'

Fwd 3: 5'-TTCCG ACCGGTCTGA TCMTTSAWAT TSCGGAWGGT TMTGMGGSGC  
15 AGRTGCKGCC GCGTAGCGGC CTGG-3'

Fwd 4: 5'-TTTTGATCGA TRSCGATTAT CRGGGCSAAS TGAWGRTTAK  
CSTGGTGAAC CKGGGCMASG AWGAAKTTRY GATTSAGCSG GGCGAACGTA  
TTGCGCAG-3'

20

Fwd 5: 5'-CGTGGCGAA GCGGGCTTTG GCTCTASCGG CASAMAGTAA  
TGAGGATCCG AATTCTT-3'

Rev A: 5'-

25 GGTCGCATAGSTCGGCAGCGGWAAWTCTTKGSCATGASGCRGACGCWKAATTTT  
CASA WYAAYTTTKYTCATATGAAGCTTGGTACCAA-3'

Rev B: 5'-

GATCAGACCGGTCGGAAYCAGCRYCSTWTCAMMCGGCKYAAKTTYCASCOSWT  
30 YSTYAAKGSMCGCACGCAGATCCAGGCCTG-3'

Rev C: 5'-

TTTTATCGATCRKGCCCRSCGCGTTCAGCASCRTAWGCCMTKTTTCCAGSCCAG  
GCCGCTACGCGGC-3'

5 Rev D: 5'-

TAGAGCCAAAGCCGCCTTCGCCACGMTSGGTCTGAGAAAMWTCTTCCACCWSA  
WYAAMTTCCRCCYGCWSCACCGGCRCAAWAAYCAKCTGCGCAATACGTTTCGCC  
C-3'

10 Rev E: 5'-AAGAATTCGGATCCTCATTACT-3'

[0124] The two resulting fragments are sub-libraries that can now be combined using classical molecular biology techniques. For example, the Fwd1/RevA/Fwd2/RevB/Fwd3/RevC fragment (the amino-encoding half) can be cloned using *NdeI* and *ClaI*. The Fwd4/RevD/Fwd5/RevE fragment (the carboxyl half) can be cloned using *ClaI* and *BamHI*. The fragments can be cloned separately than combined to form a full-length hybrid library. Alternatively, the fragments can be combined in a single step in a three-fragment forced cloning ligation.

[0125] If the vector used in the cloning is an expression vector such as pET11c, the protein can be expressed from the T7 promoter (Studier, *et al.*, *Methods in Enzymology* 185:60-89, 1990) and protein can be isolated and assayed for the desired characteristic. In this example, a thermal stable parent protein was "mixed" with a mesophilic homologue. One skilled in the art can purify these proteins (Hoffmann, *et al.*, *Eur. J. Biochem.* 164, 45-51, 1987) and assay them for their temperature optima. Differences in the sequences between proteins with different temperature optima will lead to a better understanding of the factors important in protein stabilization at high temperatures.

#### Example 2. Generation of hybrid polymerase proteins

[0126] Those skilled in the art will recognize that this example represents a much more complex application of the present invention than Example 1. Pfu polymerase is a commercially available (Stratagene, La Jolla, CA) family B DNA polymerase isolated from *Pyrococcus furiosus*. Deep Vent<sup>®</sup> is a commercially available (New England Biolabs,

Beverly, MA) family B DNA polymerase isolated from *Pyrococcus* sp. GB-D. Being 775 amino acids in length, these proteins are twice as large as a typical protein and five times as large as Dut. They share a variety of activities including DNA binding, nucleotide binding, nucleotide addition, pyrophosphorolysis, and 3' to 5' exonuclease (proofreading) activities.

5 The invention can be applied to any one of the activities encoded by these large proteins by being applied to one domain of the protein. In this example, the invention was applied to each of the different enzymatic activities, by making a hybrid library for the entire protein. Thus, this example represents at least two independent tests of the method, for the two activities assayed (polymerase activity and proofreading exonuclease activity).

10 [0127] The amino acid sequences differ from one another at 115 locations. The sequences are 85% identical over the complete sequence. One 18-amino-acid-region is only 56% identical. Hybrid Deep Vent<sup>®</sup>/Pfu proteins were produced by creating a collection of oligonucleotides that encodes a blend of sequences from the two parents and then assembling the oligonucleotides in a library of full-length polymerase proteins.

15 [0128] The protein sequence of Pfu polymerase and Deep Vent polymerase were aligned. A BlastP alignment is shown in Figure 2. As stated, the alignment found 115 differences between the Pfu and Deep Vent amino acid sequences. An *E. coli* codon usage table was then used to compare the various codons that can encode the amino acids and deduce an minimal encoding sequence. Figure 8 shows the minimal encoding sequence used to  
20 generate oligonucleotides encoding a Pfu/ Deep Vent<sup>®</sup> Hybrid DNA polymerase.

[0129] In many instances, a single nucleic acid degeneracy could encode both amino acids. For example, the parent proteins differ at amino acid position 15 where Pfu has a valine (Val) and Deep Vent an isoleucine (Ile). It is possible to encode Val using GTT and Ile using ATT. The oligonucleotide synthesis machine was therefore programmed to produce product with  
25 half G and half A at nucleotide position 43 of the protein-coding DNA. Thus, a codon with a RTT where either a G or an A is introduced into the first nucleotide position of the codon will provide a pool of oligonucleotides, some of which have a GTT at that position; the others of which have an ATT at that position.

[0130] In the alignment of Pfu and Deep Vent, 98 of the 115 differences could be simply  
30 incorporated into the library by introducing a single degeneracy at one nucleotide residue of the codon that encoded the different amino acids.

[0131] The remaining 17 differences required two nucleotides to be changed in order to encode the two parental sequences. These changes forced the possibility that two non-parental amino acid sequences would exist in the resulting library. An example of this is residue 72, at which Pfu has a glutamate (Glu) and the Deep Vent has an arginine (Arg). Glu is encoded by GAR and Arg by CGN or AGR. The minimal encoding sequence (A/G)(A/G)G was selected to potentially encode the parent sequences at position 214 through 216 of the hybrid protein-coding region. This combination will also generate nucleotides encoding glycine (GGG) and lysine (AAG). This situation was determined to be tolerable even though glycine is not similar to either parental amino acid because such situations were rare relative to the size of the protein.

[0132] Incorporation of a potential stop codon at amino acid residue 758 (nucleic acid residues 2272 and 2273) was also deemed to be tolerable. This stop codon made 1/4 of the library useless. Amino acid residue 566 (nucleotides 1696 through 1698) was made a lysine by mistake (Figure 8); it should have contained a nucleotide degeneracy that encoded lysine or aspartic acid. Figure 8 shows the minimal encoding sequence used to generate oligonucleotides encoding a Pfu/ Deep Vent® Hybrid DNA polymerase as explained in example 2. The degenerate nucleotides are in parenthesis. The amino acid sequences that differ between the parent proteins (the "mismatches") are indicated. Non-parental amino acids are indicated in bold. Examples mentioned in the text are numbered.

[0133] For each strand of the minimal encoding sequence, a set of degenerate oligonucleotides of approximately 100 bases in length, and separated by gaps of 40 bases, was synthesized. The oligonucleotide sequences on the two strands were arranged so that the oligonucleotides from the first strand spanned the gaps on the second strand and overlapped the oligonucleotides of the second strand by 30 bases (Figure 3). This oligonucleotide set was used in assembly PCR as follows. Overlapping oligonucleotides were paired, annealed to each other, and extended using a thermostable high fidelity polymerase. High concentrations of oligonucleotide and a minimal number of thermal cycles (no more than 5) were used. The products of the first cycle were double-stranded fragments of approximately 170 base pairs in length. These fragments were band-purified from a gel and used for the next cycle of pairing and primer extension to generate a new double-stranded fragment of about 310 base pairs in length. This cycle was repeated until the entire sequence was obtained as a collection of fragments of about 500 bases in length. At this point, particular fragments were selected and sequenced to assess the integrity of the procedure. It was found

that the oligonucleotides purchased were of low quality, resulting in excessive unintended mutations. A number of segments containing no unintended mutations were chosen and used to assemble full-length genes using restriction sites that had been incorporated at the ends of each fragment and conventional molecular biology techniques. Four full-length clones were assembled and the encoded proteins were expressed in pET11 (Novogene, Madison, Wi). Expression by all four clones was confirmed by SDS-PAGE. These clones were names Hyb1 to Hyb4.

[0134] A second collection of libraries was constructed on a custom basis by Blue Heron Biotechnology (Bothell, Washington) using "Genemaker" technology. The complete coding sequence was delivered as four fragment libraries that could be assembled into a full-length hybrid genes. Two full-length assembled clones were obtained and sequenced to verify validity of the library. These clones were named Phy1 and Phy2. Clones from this library contained only proper hybrid sequences including the degeneracies at position 566 (lysine/aspartic acid) and 758 (tyrosine/tryptophan) discussed earlier. The full-length sequences were cloned into expression vectors and protein of the expected size were produced.

[0135] Hybrid polymerase protein was expressed and purified from each of the six clones from the two libraries. Purification was performed as follows.

#### *Purification of hybrid polymerases*

[0136] This section describes methodology for isolating a hybrid polymerase. Following induction of expression in *E. coli*, the cells were centrifuged and the pellets stored at -20°C to -80°C. One milliliter of Buffer A (Buffer: 50 mM Tris (8.0); 50 mM Dextrose; 1 mM EDTA) was added for every 100 ml of starting culture and the cells were lysed with 4 mg/ml of powdered lysozyme at 72°C. MgCl<sub>2</sub> and CaCl<sub>2</sub> were added to a concentration of 2 mM, followed by the addition of 1 unit/ml of DNase I. The sample was shaken slowly for 10 min at room temperature. One ml of Buffer B (10 mM Tris (8.0); 50 mM KCl; 1 mM EDTA; 0.5% Tween 20; 0.5% NP40) was added per 100 mls starting culture and the sample then shaken slowly at room temperature for 15 min. The sample was transferred to a centrifuge tube and incubated at 72°C for 1 hour followed by centrifugation at 4000 x g at 4°C for 15 min. The supernatant was collected and 0.476 gm/ml of (NH<sub>4</sub>)<sub>2</sub>SO<sub>4</sub> was added and the sample was mixed slowly at 4°C for 1 hour and then centrifuged at 15,000 x g at 4°C for 15 min.

[0137] The pellet was resuspended in, and dialyzed against HiTrap Q 'A' Buffer (20 mM Tris (7.9); 50 mM NaCl; 5 mM  $\beta$ -mercaptoethanol). The suspension was then loaded onto a ÄKTAprime HiTrap Q chromatography column (Amersham Biosciences) equilibrated and run using method #2 per the manufacturers instructions using HiTrap Q buffers 'A' and 'B' ('A' buffer with 1 M NaCl). Fractions containing the polymerase were combined and dialyzed against P-11 Loading Buffer (20 mM Tris (7.9); 50 mM NaCl). The sample was bound to a liquid chromatography column of P-11 resin (Amersham Biosciences), washed with P-11 Buffer 'B' (20 mM Tris (7.9); 150 mM NaCl), then eluted using P-11 Elution Buffer (20 mM Tris (7.9); 400 mM NaCl). The eluted fractions were dialyzed against HiTrap SP 'A' buffer (20 mM Tris (6.8); 50 mM NaCl; 5 mM  $\beta$ -mercaptoethanol) then injected onto a ÄKTAprime HiTrap SP chromatography column equilibrated and run using method #2 per the manufacturers instructions using HiTrap SP 'A' and 'B' Buffer ('A' buffer with 1 M NaCl). Fractions containing PhS1 were concentrated using a YM-30 Centricon protein concentrator (Millipore). The sample was then dialyzed against buffer containing 50 mM Tris (pH 8.2); 0.1 mM EDTA; 1mM DTT; 0.1% NP40; 0.1% Tween 20. The final volume was then measured and 1.47X 85% glycerol, and 0.015X 10% NP-40 and 10% Tween 20 added. The sample was stored at  $-20^{\circ}\text{C}$ .

[0138] Of the six hybrid polymerase proteins generated from the two libraries, all had DNA polymerase activity.

[0139] Sso7d fusion polymerases (*see, e.g.*, WO0192501) were prepared using some of the hybrid polymerase proteins and compared to the parental Pfu polymerase with and without Sso7d (designated as "Pfu" and "Pfs", respectively) in exonuclease assays and extension assays. Sso7d fusions of Hyb clones are designated HyS; Sso7d fusions of the Phy clones are designated PhS. The most thoroughly studied hybrid protein was PhS1.

[0140] To measure exonuclease activity, a 45 base long primer with the following sequence was synthesized: 5'-FAM-TTTTTTGAGGTGTGTCCTACACAGCGGAGTGTAGGACACACCTCT\* 3', wherein T\* is an amino-link dT with the quencher, DAB (dabcyl) attached. The sequence forms a 16 base pair stem loop structure with a T:T\* mismatch at the quencher-labeled base. The 5' unbase-paired poly T sequence keeps FAM (6 carboxy-fluorescein) in close proximity to the quenching dye so the FAM, if excited, it will not fluoresce.

[0141] The oligonucleotide was combined with buffer and the enzyme and incubated in a real time detection instrument, the DNA Engine Opticon System (MJ Research, Inc.). This instrument excites the FAM and detects any fluorescence if present. In the absence of 3' to 5' exonuclease activity, there is only background fluorescence because FAM is quenched by DAB. However if the enzyme does have 3' to 5' exonuclease activity, the T:T\* mismatch is recognized and the 3'-T\* is removed. The DAB is released and will no longer quench the FAM fluorescence. The Opticon System will detect the increase in fluorescence with increasing time (readings were taken every 10 sec at 65°C). The rate of fluorescence increase directly reflects the amount of 3' to 5' exonuclease activity. An increase in fluorescence greater than control levels shows that the enzyme has 3' to 5' exonuclease activity. The results (Figure 9) of this analysis are discussed below.

[0142] Figure 10 shows results of a comparison of a hybrid and a parent polymerase in extension assays. Even with excess enzyme (80 U/ml), Pfu could not amplify any amplicon longer than 2 kb. An Sso7d fusion to Pfu polymerase (PfS) amplified a 10 kb fragment given a 1 min extension time. PhS1 amplified a 15 kb fragment (arrow) in 80 mM KCl with a 1 minute extension time. Further, PhS1 was also able to perform long PCR under a variety of salt conditions.

#### *Characterization of additional hybrid polymerases*

[0143] Five additional hybrid clones were isolated from the second library directly as Sso7d fusions and were designated PhS3 to PhS7. The polymerases were tested for polymerase and exonuclease activity. Table 1 summarizes characteristics of the various hybrid proteins analyzed in this example. PhS2 has two mutations at sites other than a target site. PhS3 is truncated due to an early stop codon. PhS4 has one deletion and one mutation. The "Hyb" and "HyS" polymerases also comprise mutations at positions other than the target sites, probably due to faulty oligonucleotide synthesis.

Table 1

Pol	Activity	Full-length	KCL Opt	Temp. Stab.	Processivity	Number Pfu parent residues	Number D. vent parent residues	Relative specific activity
PhS 1	Yes	Yes	80-100 mM	3 hr, 97.5	26-30	55	60	1.5
PhS2	Yes	Yes	160-180 mM	3 hr+, 97.5	24-28	64	51	4
PhS 3	No	No	N/A	N/A	N/A	N/A	N/A	n.d.
PhS 4	No	No; minus one Pfu/DV amino acid	N/A	N/A	N/A	56	58	n.d.
PhS 5	Yes	Yes	40-80 mM	3 hr, 97.5	nd	52	63	1
PhS 6	No	No	N/A	N/A	N/A	55	60	n.d.
PhS 7	Yes	Yes	40-80 mM	3 hr, 97.5	nd	54	61	2
Hyb1	Yes	Yes	nd	10 min*	2-4 nt	59	46	n.d.
Hys1	Yes	Yes	90-100mM	8-14 min*	11 nt	59	46	2
Hyb2**	Yes	No	nd	n.d.	n.d.	50	53	n.d.
Hyb3**	Yes	No	nd	n.d.	n.d.	51	47	n.d.
Hys4	Yes	Yes	80-90 mM	< 1min*	n.d.	51	50	n.d.

All polymerases designated "PhS" are Sso7d fusions.

"Hys1" is Hyb1 with Sso7d at the C-terminus.

"Hys4" has Sso7d at the C-terminus.

5

[0144] The exonuclease activity of various hybrid polymerases was also evaluated as described above. The polymerase to 3'-exonuclease ratios for several commercially available enzymes, including the parental proteins and isolates from the hybrid library, were compared. DyNAzyme EXT, an enzyme used in long accurate PCR, is a blend of a Family B polymerase with 3' to 5' exonuclease activity, and a Family A polymerase that lacks any proofreading activity. Too much exonuclease activity is detrimental because it digests primers instead of extending them. Pfu and Deep Vent are the parental Family B polymerases which both have high exonuclease activity. Pfs (a Pfu-Sso7d fusion enzyme) has increased polymerase activity. HyS1, PhS1, PhS2, PhS5, and PhS7 are isolates from the hybrid libraries. Surprisingly, the results (Figure 9) show that the hybrid proteins vary greatly in their polymerase to exonuclease activities, both relative to the parent proteins and each other. PhS1 has a polymerase to exonuclease activity ratio approaching that of the enzyme blend.

10

15

[0145] A comparison of the sequences of the parent and hybrid proteins is presented in Figure 11.

[0146] These results show that multiple polymerase hybrid isolates from two different libraries were active. Furthermore, the example shows that the method also allows for  
5 generating hybrids for different domains, *i.e.*, polymerase activity domain vs. exonuclease activity domain. Clearly, the invention could be applied to proteins with very divergent activities

[0147] It is understood that the examples and embodiments described herein are for  
10 illustrative purposes only and that various modifications or changes in light thereof will be suggested to persons skilled in the art and are to be included within the spirit and purview of this application and scope of the appended claims.

[0148] All publications, patents, and patent applications cited herein are hereby incorporated by reference in their entirety for all purposes.